Malicious User Detection Based on Low-Rank Matrix Completion in Wideband Spectrum Sensing

Zhijin Qin, *Member, IEEE,* Yue Gao, *Senior Member, IEEE* and Mark D. Plumbley, *Fellow, IEEE*

***Abstract*—In cognitive radio networks, cooperative spectrum sensing (CSS) has been a promising approach to improve sensing performance by utilizing spatial diversity of participating secondary users (SUs). In current CSS networks, all cooperative SUs are assumed to be honest and genuine. However, the presence of malicious users sending out dishonest data can severely degrade the performance of CSS networks. In this paper, a framework with high detection accuracy and low costs of data acquisition at SUs is developed, with the purpose of mitigating the influences of malicious users. More specifically, a low-rank matrix completion based malicious user detection framework is proposed. In the proposed framework, in order to avoid requiring any prior information about the CSS network, a rank estimation algorithm and an estimation strategy for the number of corrupted channels are proposed. Numerical results show that the proposed malicious user detection framework achieves high detection accuracy with lower data acquisition costs in comparison with the conventional approach. After being validated by simulations, the proposed malicious user detection framework is tested on the real-world signals over TV white space spectrum.**

***Index Terms*—cooperative spectrum sensing, low-rank matrix completion, malicious user detection, TV white space.**

I. Introduction

SPECTRUM sensing, a promising solution to identify potential spectral holes, is one of the most challenging tasks in cognitive radio (CR) networks [1]. Cooperative spectrum sensing (CSS) is an effective approach to offer significant performance gain in detecting spectrum holes, by exploiting the spatial diversity of collaborative secondary users (SUs) [2, 3]. However, due to the openness of low-layer protocol stacks, CSS networks are vulnerable to attacks from spectrum sensing data falsification (SSDF) [4]. This characteristic of CSS networks blocks the application of CR techniques in large-scale networks.

In CSS networks, SUs that launch SSDF attackers are called *malicious users*. The main goals of malicious attacks come from two aspects: 1) decreasing detection probability for disturbing the normal operation of PUs; and 2) increasing false alarm probability to deprive access opportunities for honest SUs [5]. In decentralized CSS networks, sensing results are exchanged between neighboring SUs to improve the network reliability to link failure. However, this characteristic makes decentralized CSS more vulnerable to malicious attacks [6], as the observations at honest SUs are also available to malicious users during the information exchanging and convergence process. Furthermore, corrupted data can be integrated into the decisions of honest neighbor SUs, which eventually brings significant performance degradation to the whole CSS network [7]. In centralized CSS networks, all SUs report their local sensing data to a fusion center (FC), where the final decision on spectrum occupancy is made. By doing so, all participating SUs, including malicious users, can only obtain the spectrum occupancy knowledge from the FC. Thus, observations at honest SUs in the CSS network would not leak to malicious users directly. However, as the corrupted data are still considered in the decision making process, the existence of malicious users may lead to incorrect decisions at the FC. Generally, regardless of the types of malicious attack and CSS network, malicious users pose significant challenges to the security in CSS networks. Therefore, accurate malicious user detection is essential to guarantee the security of CSS networks.

Along with improving the security of CSS networks through malicious user detection, another key challenge in attempting to build CSS networks comes from the need to reduce data acquisition costs at SUs. Due to the well-known Nyquist sampling theorem, sampling rates should be no less than twice the signal bandwidth. However, as the SUs are normally energy-constrained, their sensing capabilities are normally limited as well. Thus, it is difficult to achieve such high sampling rates for wideband spectrum sensing at energy-constrained SUs. Additionally, as the spectrum is normally underutilized in reality [8, 9], the signal spectrum exhibits a sparsity property [10] in the frequency domain. It is further noted that this sparsity property can be transformed into a low-rank property of the matrix constructed by spectral signals received at spatially distributed SUs [11], since nearby locations or adjacent channels share the similar spectrum occupancies [12]. Matrix completion (MC) techniques [13] can be applied to recover the complete matrix with only partial of the items observable. Specifically, by invoking MC technique at the FC, SUs in a CSS network need to sense fewer channels, as the unsensed channels can be reconstructed from the sensed channels based on the low-rank property. As a result, the data acquisition costs are significantly reduced at SUs.

A. Related Work

So far, malicious user detection has been widely researched for enhancing the security of CSS networks [14–23]. Specifically, the performance of CSS networks with single and

Zhijin Qin is with Lancaster University, Lancaster, United Kingdom. (email: zhijin.qin@lancaster.ac.uk)

Yue Gao are with Queen Mary University of London, London E1 4NS, United Kingdom. (email: yue.gao@qmul.ac.uk)

Mark D. Plumbley is with the Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey, Guildford, Surrey GU2 7XH, United Kingdom. (email:m.plumbley@surrey.ac.uk)

multiple malicious users were investigated by Wang *et al.* [14]. Here, the suspicious level of each SU as well as consistency values were calculated based on historical reports from SUs, in order to reduce the influences of malicious users. Min *et al.* [19] proposed to safeguard the detection process by checking the consistency among sensing reports and and removing the abnormal sensing reports. Chen *et al.* [15] proposed a reputation-based mechanism to defend against malicious attacks. However, these historical-data-based algorithms take a long time to build a reliable reputation. Kaligineedi *et al.* [16] proposed a robust outlier detection method to identify malicious users that keep sending high values regardless of the spectrum occupancies, by utilizing outlier factors and spatial information about SUs. The shadowing correlation between SUs were exploited in [17] to detect reports manipulated by malicious users. Kalamkar *et al.* [18] proposed an outlier detection scheme to detect malicious users which send true or false power values randomly. Penna *et al.* [20] proposed a Bayesian method to detect and counteract attacks in CSS networks by considering the attack probability at each participating SU. Malicious users are assumed to send the false reports and the achieved performance is better than schemes based on exclusion of the malicious users. Duan *et al.* [21] proposed two attack-prevention mechanisms with direct and indirect punishments. Furthermore, some work has been carried out on attack detection from intelligent malicious users. Li *et al.* [22] proposed an abnormality-detection approach for secure CSS networks, in which the intelligent attack strategy adopted by malicious users is unknown. Wang *et al.* [23] designed an incentive compatible mechanism to thwart malicious behaviors. By doing so, the proposed approach was more practical to be implemented in CR networks.

As well as the existing work on malicious user detection, MC-based CSS networks have been studied [12, 24–29], with the purpose of alleviating the costs of data acquisition at SUs. Meng *et al.* [24] introduced the concept of MC to CSS networks. In [24], each SU linearly combines the information on multiple channels at sub-Nyquist sampling rates. Subsequently, each SU sends a small number of such linear combinations to an FC for MC. Li [12] applied a belief propagation framework to MC to make the spectrum reconstructing more implementable and efficient in the wideband CSS networks. Qin *et al.* [25] proposed a robust wideband spectrum sensing algorithm for TV white space (TVWS), with sub-Nyquist sampling performed at each SU in the considered CSS networks. Once the compressed measurements are sent to the FC, nuclear norm minimization is adopted to solve the low-rank MC problem [26, 27]. By doing so, the costs of data acquisition at SUs are reduced significantly. For spectrum allocation over TVWS, the geo-location database approach can provide the spectrum occupancy information for secondary usage. By utilizing the information from the geo-location database, a prior information assisted wideband spectrum sensing algorithm was proposed in [28], with the purpose of improving the detection performance and reducing the costs of data acquisition at SUs. Additionally, Qin *et al.* [29] proposed to remove the corrupted samples at the FC with sub-Nyquist sampling rates at SUs. However, rank of the matrix at the FC and malicious user number are assumed to be known in advance in the considered CSS networks. The two strong assumptions make the malicious user detection algorithm difficult to be implemented in dynamic CSS networks. Moreover, the transmission cost from SUs to the FC is high as all the collected measurements are transmitted to a FC, which is inefficient or even unaffordable for large-scale networks.

B. Motivation and Contribution

The aforementioned methods [14–23] have played a vital role and laid a solid foundation to foster new strategies for malicious user detection. However, many of these methods are trust-based, which uses historical information on malicious users' behaviours. In practice, reliable reputation information is not always available, since well-established historical statistics may be too expensive or even unrealistic in a fast-changing CR environment. Additionally, intelligent malicious users sending random false values, but very close to the true ones, are more challenging than the types of malicious users that always send very high or very low values [16, 18]. Motivated by these findings, a malicious user detection method, dealing with malicious users sending values that are random but within the range of true power values reported by cooperative SUs, is desirable for secure CSS networks. Another motivation of our work comes from the need to reduce data acquisition costs at SUs without loss of any information. The aforementioned MC-based CSS methods [12, 24–26, 28, 29] focus on reducing the costs of data acquisition at SUs without considering any security issues or require high data transmission cost in CSS networks. Therefore, a malicious user detection algorithm with energy efficiency at SUs is strongly needed, but challenging.

To the best of our knowledge, this is the first work which invokes the MC technique to achieve malicious user detection without prior information about the CSS networks. The primary contributions of this paper are summarized as follows:

- A malicious users detection framework is proposed based on low-rank MC. In the proposed framework, only part of the whole spectrum is sensed without degrading the recovery performance, as the unsensed channels can be recovered from the sensed channels at the FC by invoking a MC technique. Consequently, the data acquisition costs at SUs are reduced in comparison with the CSS networks without MC.
- At the FC, channels which are sensed but corrupted by malicious users are removed during the MC process, by utilizing the adaptive outlier pursuit (AOP) algorithm [30, 31]. Therefore, malicious users are removed before making final decisions on spectrum occupancies.
- A rank estimation algorithm is proposed to provide the rank as one of the inputs for the proposed malicious user detection algorithm. By doing so, the proposed framework does not require any prior information about spectrum sparsity for malicious user detection.
- An estimation strategy for the number of channels corrupted by malicious users is proposed. With the new strategy, the estimated number of corrupted channels can

be used as one of the inputs for the proposed malicious user detection algorithm to ensure that the proposed framework is completely blind to the CSS network.
- The proposed framework is tested on real-world signals over TVWS after being validated by simulated signals. Numerical results show that the proposed malicious user detection framework achieves high detection accuracy with lower data acquisition costs at SUs, in comparison with the conventional CSS algorithms.

C. Organizations

The rest of this paper is organized as follows. Section II describes the CSS network model with malicious users. Section III presents our proposed malicious user detection framework along with the proposed rank estimation algorithm and the estimation strategy for the number of corrupted channels. Section IV shows the numerical analyses of the proposed malicious user detection framework on both simulated and real-world signals. Section V concludes this paper.

II. Network Model of Cooperative Spectrum Sensing with Malicious Users

A. Network Description

We take a typical CSS scenario as the considered network model, as shown in Fig. 1(a). It is assumed that the whole spectrum of interest with bandwidth $\mathcal{D}$ can be divided into $\mathcal{I}$ channels. A channel is either occupied by a PU or unoccupied. There is no overlap between different channels. The number of occupied channels K is assumed to be much less than the total number of channels, i.e. $K \ll \mathcal{I}$. Each channel is sensed by a set of SUs which are spatially randomly distributed at different J locations. At an arbitrary location indexed by j $(1 \leq j \leq J)$, it is assumed that a set of B_j $(1 \leq B_j \leq \mathcal{I})$ SUs are deployed to sense the spectrum of interest. Within each set of SUs, SUs are indexed by b $(1 \leq b \leq B_j)$.

In a conventional CSS network, the whole spectrum is sensed by one SU at each location, which results in $B_j = 1$ and J SUs deployed in the CSS network in total. However, high sampling rates are challenging for SUs to achieve wideband spectrum sensing, as the SUs are normally energy-constrained with limited sensing capabilities. In this paper, we propose to deploy a set of SUs at each location (i.e. $B_j > 1$), where each SU is adjusted to sense different channels among the whole spectrum at Nyquist rates. This can be achieved by equipping each SU with a bandpass filter. Additionally, the whole spectrum is not fully sensed by the deployed SU set at each location. As shown in Fig. 1(a), only part of the whole spectrum are sensed by B_j SUs at the j-th location, and each of the SU is labelled as SU_{bj}. In other words, at each location, each channel is sensed by one SU at most, and some of the channels are not sensed by any SU. Consequently, the costs of data acquisition at SUs can be reduced significantly, in comparison with the case that each SU senses the whole spectrum.

After sampling is performed, each SU calculates the power values of the sensed channels, and then sends this information to an FC to contribute to the final decisions on spectrum occupancies. It is further noticed that some of the SUs experience deep fading or shadowing. They would send very low power values to the FC in a CSS network regardless of the spectrum occupancies, which are labelled as blocks with '+' in Fig. 1(a) and Fig. 1(b). The transmit signal has a sparsity property in the frequency domain [10] and the nearby locations are assumed to share the similar spectrum occupancies [12], so the matrix constructed by the received signals at different locations exhibits a low-rank property [11]. Fig. 1(b) illustrates the transformation of the sparsity property of transmit signals into the low-rank property of the matrix at the FC, where the matrix is constructed by signal powers of the sensed channels received at different locations. In such a CSS network, we need to reconstruct the unsensed channels from the sensed channels by a low-rank MC technique.

In the case of a sensing malfunction, some SUs in the CSS network send corrupted power values to the FC, labelled as the blocks with 'X' in Fig. 1(a) and Fig. 1(b). Malicious users appear randomly in the considered CSS network to corrupt a random number of channels at random locations. Malicious users that keep sending high power values or low power values are easily detected. However, malicious users sending values that are random but very close to the true values are much more difficult to detect, which is the case we consider in this paper. We propose to remove these malicious users during the MC process at the FC. Otherwise, recovery performance at the FC would be degraded significantly as the corrupted power values are used for the MC.

B. Signal Processing Model

Let us define $s_i(t) \in \mathcal{C}^{N\times 1}$ to be the signal transmitted over the i-th channel, where N refers to the number of samples at the Nyquist sampling rate. It is assumed that only the primary signals are transmitted in the spectrum of interest. Additionally, $r_{ij}(t)$ refers to the signals over the i-th channel received by the SU at the j-th location, which is given by

$$r_{ij}(t) = d_j^{-\chi/2} h_{ij} s_i(t), \tag{1}$$

where d_j refers to the distance from PUs to the receiver at the j-th location, χ is defined as the propagation loss factor, and h_{ij} follows independent Rayleigh distribution which is assumed to be time-invariant within one sensing period.

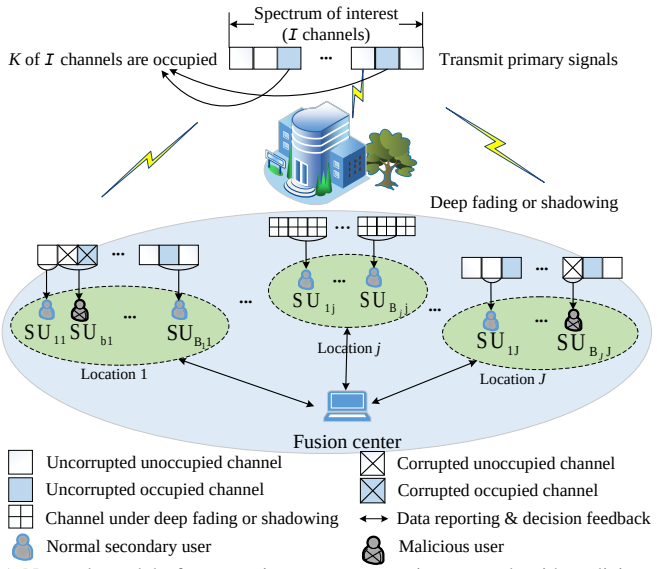
Once signals over the i-th channel $r_{ij}(t)$ are received at an SU at the j-th location, the power of the sensed channel p_{ij} can be calculated as

$$p_{ij} = \frac{1}{N}\sum_{n=1}^{N} |r_{ij}[n]|^2, \tag{2}$$

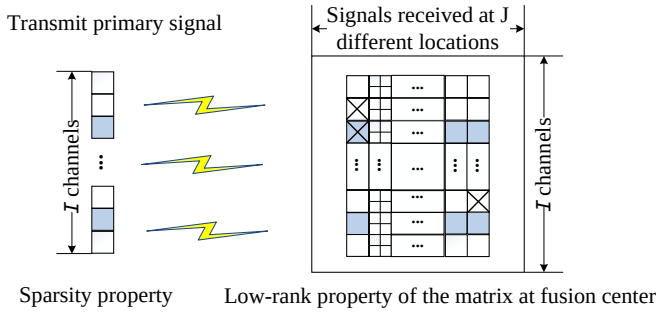
where $r_{ij}[n]$ refers to the discrete sampling of $r_{ij}(t)$. The complete matrix $\mathbf{P^{\Omega}}$ constructed by signals received at J different locations can be illustrated as

$$\mathbf{P^{\Omega}} = \begin{bmatrix} p_{1,1} & \cdots & p_{1,J} \\ \vdots & \ddots & \vdots \\ p_{\mathcal{I},1} & \cdots & p_{\mathcal{I},J} \end{bmatrix}_{\mathcal{I}\times J} \tag{3}$$

where the i-th row of $\mathbf{P^{\Omega}}$ represents the power values of the i-th channel sensed by SUs located at the different J locations.



(a) Network model of cooperative spectrum sensing network with malicious users.



(b) Transformation of sparsity property into low-rank property.

Fig. 1: Network model

The the j-th column of $\mathbf{P}^{\mathbf{\Omega}}$ refers to the power values of different channels sensed by the B_j SUs at the j-th location. Here, $\mathbf{\Omega}$ is defined as the index set of complete matrix $\mathbf{P}^{\mathbf{\Omega}}$.

As the whole spectrum of interest is not fully sensed by the set of SUs at each location in the considered network, the matrix constructed by the power values available at the FC is incomplete and labelled as $\mathbf{P}^{\mathbf{E}}$. The item $p_{ij}^{\mathbf{E}}$ in $\mathbf{P}^{\mathbf{E}}$ can be expressed as

$$p_{ij}^{\mathbf{E}} = \begin{cases} p_{ij} & \text{if } (i,j) \in \mathbf{E} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $\mathbf{E}$, a subset of $\mathbf{\Omega}$, denotes the index set for sensed channels at different locations. The elements in $\mathbf{E}$ and the number of elements in $\mathbf{E}$ are both random. We denote the average number of elements in $\mathbf{E}$ as $\mathcal{E}\{|\mathbf{E}|\}$ and $\gamma = \frac{\mathcal{E}\{|\mathbf{E}|\}}{|\mathbf{\Omega}|}$. Here, $|\mathbf{\Omega}|$ refer to the number of items in $\mathbf{E}$ and $\mathbf{\Omega}$, respectively. Specifically, $p_{ij}^{\mathbf{E}}$ can also be expressed as

$$p_{ij}^{\mathbf{E}} = \begin{cases} p_{ij} & \text{w.p. } \gamma \\ 0 & \text{w.p. } 1-\gamma. \end{cases} \tag{5}$$

We define $\mathbf{O}$ as a subset of $\mathbf{E}$ to denote the index set for the sensed channels from honest SUs at different locations. We assume that a sensed channel is from a honest SU with probability κ. Then it is obvious that $\kappa = \mathcal{E}\left\{\frac{|\mathbf{O}|}{|\mathbf{E}|}\right\}$. If $SU_{i,j}$ decides to corrupt the measured channel power p_{ij}, it will generate a corrupted power value. Then with the existence of malicious users, the power of an arbitrary sensed channel can be expressed as

$$\tilde{p}_{ij} = \begin{cases} \tau\left(p_{\max}^{\mathbf{E}} - p_{\min}^{\mathbf{E}}\right) + p_{\min}^{\mathbf{E}} & \text{w.p. } 1-\kappa \\ p_{ij} & \text{w.p. } \kappa \end{cases} \tag{6}$$

where $\tau \sim U(0,1)$ follows a standard uniform distribution with minimum 0 and maximum 1, $p_{\min}^{\mathbf{E}}$ and $p_{\max}^{\mathbf{E}}$ refer to the minimum and maximum power values of the sensed channels, respectively. More specifically, $p_{\min}^{\mathbf{E}}$ refers to the power value of the vacant channel, and $p_{\max}^{\mathbf{E}}$ refers to power value of the occupied channel, which can be known at SUs according to historical information.

With malicious users appearing in the CSS network, the incomplete matrix $\mathbf{P}^{\mathbf{E}}$ becomes sparsely corrupted, labelled as $\mathbf{P}^{\mathbf{EC}}$. Then the item $p_{ij}^{\mathbf{EC}}$ in $\mathbf{P}^{\mathbf{EC}}$ can be expressed as

$$p_{ij}^{\mathbf{EC}} = \begin{cases} \tilde{p}_{ij} & \text{w.p. } \gamma \\ 0 & \text{w.p. } 1-\gamma. \end{cases} \tag{7}$$

III. Proposed Malicious User Detection Framework

In order to enhance the security of the CSS network, a malicious user detection framework based on low-rank MC is proposed. Based on the network model described in Section II-A, each SU is proposed to sense only part of the spectrum rather than the whole spectrum, in order to reduce costs of data acquisition at each SU. Additionally, we propose to remove the power values of corrupted channels at the FC by invoking the AOP algorithm [30]. It is further noted that the rank of the matrix at the FC and the number of channels corrupted by malicious users is normally unknown in reality, but are required by the proposed malicious user detection algorithm with AOP. To make the proposed malicious user detection framework completely blind, a rank estimation algorithm and an estimation strategy on the number of corrupted channels collected at the FC are proposed. As a result, the malicious user detection framework does not require any prior information about the CSS network. Once the exact matrix is obtained by the proposed framework, spectrum occupancies can be determined by a conventional energy detection method [32].

A. Malicious User Detection Algorithm Based on Matrix Completion

With malicious users appearing in a CSS network, power values of the sensed channels are partly corrupted, affecting the accuracy of recovery of sensed channels at the FC. As attacks can be launched by any malicious SU participating the CSS network, corrupted power values are assumed to be sparsely and randomly distributed among the collected values at the FC. The indices of the corrupted power values are unknown to the FC. Additionally, in order to make attacks more difficult to detect, each malicious user may distort the power of an arbitrary channel to a value that is close to the true

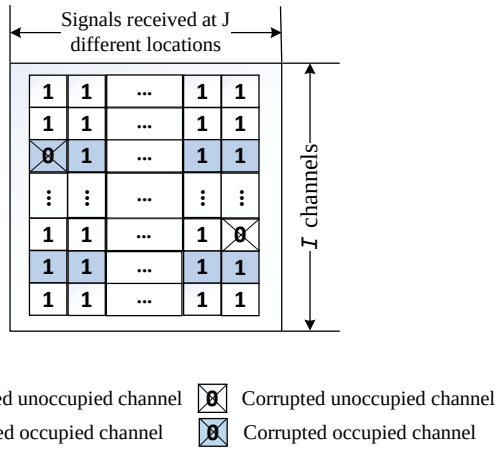


Fig. 2: Illustration for the corruption index matrix $\mathbf{\Lambda}$, with '1' for uncorrupted channels and '0' for corrupted channels received at the fusion center.

power. In such a case, the malicious user detection problem can be formulated as follows:

$$\begin{aligned}&\hat{\mathbf{P}}^{\mathbf{\Omega}}, \hat{\mathbf{\Lambda}} = \arg\min_{\mathbf{P}^{\mathbf{\Omega}}, \mathbf{\Lambda}} \tfrac{1}{2} \sum_{(i,j)\in\mathbf{\Omega}} \mathbf{\Lambda}_{ij} \left(\left(\mathbf{P}^{\mathbf{\Omega}}\right)_{ij} - p_{ij}^{\mathbf{EC}} \right)^2 \\ &\text{s.t.} \sum_{(i,j)\in\mathbf{\Omega}} (1 - \mathbf{\Lambda}_{ij}) \le L,\ \mathbf{\Lambda}_{ij} \in \{0,1\},\end{aligned} \tag{8}$$

where $\hat{\mathbf{P}}^{\mathbf{\Omega}}$ is the recovered matrix, and the number of corrupted power values collected at the FC is L, where is unknown and to be discussed in the following. Binary matrix $\mathbf{\Lambda}$ denotes the uncorrupted channels at different locations by 1, i.e. $\mathbf{\Lambda}_{ij} = 1$ if $(i,j) \in \mathbf{O}$ for channel i at the j-th location, and the corrupted channels by 0, which is illustrated in Fig. 2.

It is noted that problem (8) is non-convex, since it has both continuous and discrete variables. The following two steps can be performed to find a local optimal solution to (8):

1) Fix $\mathbf{\Lambda}$ and update $\mathbf{P}^{\mathbf{\Omega}}$. If $(i,j) \notin \mathbf{\Omega}$, $\Lambda_{ij}\left(\left(\mathbf{P}^{\mathbf{\Omega}}\right)_{ij} - p_{ij}^{\mathbf{EC}} \right)^2$ in (8) becomes zero. Therefore, $\hat{\mathbf{P}}^{\mathbf{\Omega}}$ can be obtained by solving the simplified problem

$$\hat{\mathbf{P}}^{\mathbf{\Omega}} = \arg\min_{\mathbf{P}^{\mathbf{\Omega}}} \sum_{(i,j)\in\mathbf{\Omega}} \left(\left(\mathbf{P}^{\mathbf{\Omega}}\right)_{ij} - p_{ij}^{\mathbf{EC}} \right)^2. \tag{9}$$

This problem can be easily solved by Riemannian trust-region for MC (RTRMC) [33].

2) With fixed $\mathbf{P}^{\mathbf{\Omega}}$, $\mathbf{\Lambda}$ can be updated by solving

$$\begin{aligned}&\hat{\mathbf{\Lambda}} = \arg\min_{\mathbf{\Lambda}} \sum_{(i,j)\in\mathbf{\Omega}} \mathbf{\Lambda}_{ij} \left(\left(\mathbf{P}^{\mathbf{\Omega}}\right)_{ij} - p_{ij}^{\mathbf{EC}} \right)^2 \\ &\text{s.t.} \sum_{(i,j)\in\mathbf{\Omega}} (1 - \mathbf{\Lambda}_{ij}) \le L, \mathbf{\Lambda}_{ij} \in \{0,1\}.\end{aligned} \tag{10}$$

The problem (10) is to choose $(\mathcal{I} \times J - L)$ elements with least sum from $\boldsymbol{S}_{\mathbf{\Omega}} = \left\{ \left(\left(\mathbf{P}^{\mathbf{\Omega}}\right)_{ij} - p_{ij}^{\mathbf{EC}} \right)^2, (i,j) \in \mathbf{\Omega} \right\}$. Given Γ as the L-th largest item in $\boldsymbol{S}_{\mathbf{\Omega}}$, $\mathbf{\Lambda}_{ij}$ can be updated as

$$\mathbf{\Lambda}_{ij} = \begin{cases} 1 & \text{if } (i,j) \in \mathbf{\Omega}, \left(\left(\mathbf{P}^{\mathbf{\Omega}}\right)_{ij} - p_{ij}^{\mathbf{EC}} \right)^2 < \Gamma \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

During the process of solving problem (8), the power values for those corrupted channels are removed from the observed ones at the FC to alleviate the recovery errors caused by malicious users. Once the recovered matrix $\hat{\boldsymbol{P}}^{\mathbf{\Omega}}$ is obtained at the FC, spectrum occupancies are determined by comparing the energy of each channel with a predefined threshold. Particularly, the i-th channel is determined as occupied if its energy is higher than the empirical threshold λ_d as given by

$$\lambda_d = \left(\frac{\mu}{2}\right)^2, \tag{12}$$

where $\mu = \left\| \mathrm{vec}\left(\hat{\mathbf{P}}^{\mathbf{\Omega}}\right) \right\|_1 \Big/ \left\| \mathrm{vec}\left(\hat{\mathbf{P}}^{\mathbf{\Omega}}\right) \right\|_0$ refers to the average absolute value of all the $J \times K$ nonzero elements in $\mathrm{vec}\left(\hat{\mathbf{P}}^{\mathbf{\Omega}}\right)$ [11]. Here, $\mathrm{vec}\,(\cdot)$ stacks all columns of matrix $\hat{\mathbf{P}}^{\mathbf{\Omega}}$ into a long vector. The final binary decisions $\mathbf{d} = (d_1, \ldots, d_{\mathcal{I}})$ on spectrum occupancies can be determined as

$$d_i = \left(\frac{1}{J} \sum_{j=1}^{J} \sum_{i=1}^{\mathcal{I}} |\hat{p}_{ij}| \ge \lambda_d \right),\ \forall i \tag{13}$$

where $\hat{p}_{ij}$ is the reconstructed power value of the i-th channels sensed by the SU at the location j.

The proposed malicious user detection algorithm is summarized in Algorithm 1. The maximal number of iterations for solving (8) is predefined as $I_{\max}$. At the l-th iteration, $\mathbf{\Lambda}^l$ is used to identify the corrupted power values based on the newly constructed $\left(\mathbf{P}^{\mathbf{\Omega}}\right)^l$. Then the indices for the uncorrupted power values are updated to be 1 and the other items are indexed by 0 in order to remove the corrupted values during the matrix completion process. This iterative process continues until the index set for the uncorrupted power values is same as that in the last iteration. It should be noted that malicious users that send both corrupted and uncorrupted power values are not removed completely during the iterative recovery process. Only the corrupted values are removed, which maximizes the number of uncorrupted measurements collected at the FC to achieve better recovery performance. After the complete matrix is recovered, final decisions on spectrum occupancies $\mathbf{d}$ are made according to (13). As demonstrated in Algorithm 1, the rank of the matrix K and the number of corrupted power values L are required as the inputs at the FC. However, the two kinds of prior information are normally unknown in reality. Therefore, we propose a rank estimation algorithm and an estimation strategy for the number of corrupted channels collected at the FC, to enable the proposed malicious user detection framework. The details of the proposed algorithm and strategy are introduced in the following two parts.

B. Rank Estimation Algorithm

Rank estimation can be converted into a sparsity order estimation problem, as the sparsity property of a signal can be transformed into a low-rank property of a matrix constructed by signals received by SUs at different locations [11]. With the sensed channels observed at the FC, rank of a matrix can be estimated using the following two steps:

Algorithm 1 Proposed Malicious User Detection Algorithm

Input: $\mathbf{\Omega}$, $\hat{M}$ sensed channels, $I_{\max} > 0$, $L \geq 0$, $K > 0$.
Initialization: $l = 1$, $\Lambda_{ij} = 1$ for $(i,j) \in \mathbf{\Omega}$, $\mathbf{O}^0 = \mathbf{\Omega}$.
while $l < I_{\max}$ **do**
  Update $\left(\mathbf{P}^{\mathbf{\Omega}}\right)^l$ with RTRMC as in (9);
  Update $\mathbf{\Lambda}^l$ with (10);
  Update $\mathbf{O}^l$ to be indices in $\mathbf{\Omega}$ where $\mathbf{\Lambda}^l_{ij} = 1$;
  if $\mathbf{O}^l = \mathbf{O}^{l-1}$ **then**
    $\hat{\mathbf{P}}^{\mathbf{\Omega}} = \left(\mathbf{P}^{\mathbf{\Omega}}\right)^l$;
    Break;
  end if
  Update $l = l + 1$;
end while
$\hat{p}_{ij} = \left(\hat{\mathbf{P}}^{\mathbf{\Omega}}\right)_{ij}$.
Calculate threshold λ_d.
Make final decisions $\mathbf{d}$ on spectrum occupancies by (13).
return $\mathbf{d}$.

1) The reconstructed complete matrix $\bar{\mathbf{P}}^{\mathbf{\Omega}}$ for rank order estimation can be obtained by solving

$$\bar{\mathbf{P}}^{\mathbf{\Omega}} = \arg\min_{\mathbf{P}^{\mathbf{\Omega}}} \left\|\mathrm{vec}\left(\mathbf{P}^{\mathbf{\Omega}}\right)\right\|_1 \text{ s.t. } \mathcal{A}\left(\mathbf{P}^{\mathbf{\Omega}}\right) = \mathbf{P}^{\mathbf{EC}} \tag{14}$$

where $\mathcal{A}$ is an operator from $\mathbf{\Omega}$ to $\mathbf{\Omega}/\mathbf{E}$, and the AIC sampler [34] is taken in this paper. Here, the sparsity level of $\mathrm{vec}\left(\bar{\mathbf{P}}^{\mathbf{\Omega}}\right)$ is equal to $J \times K$.

2) The estimated rank $\hat{K}$ is given by

$$\hat{K} = \sum_{i=1}^{\mathcal{I}} \left(\left| \frac{1}{J} \sum_{j=1}^{J} \bar{p}_{ij} \right| \geq \lambda_r \right) \tag{15}$$

where λ_r is a threshold for the rank estimation, and $\bar{p}_{ij}$ refers to the items in $\bar{\mathbf{P}}^{\mathbf{\Omega}}$. By applying data fusion at the FC, the power value of each channel is then calculated by averaging power values of the same channel received by spatially distributed SUs.

It has been proved that exact signal recovery can be guaranteed when the number of sensed channels is no less than $\alpha\left(K \times J\right)\log\left(\mathcal{I}/K\right)+\beta$ [35]. However, the number of sensed channels that guarantees exact rank estimation and exact signal recovery are different. When Monte Carlo simulation and statistical curve fitting techniques are adopted to find the values of the constants α and β, the following two results can be obtained with given $\mathcal{I}$, J and K:

Result 1: Successful rank estimation can be guaranteed when the number of sensed channels is not less than M_1, which is defined as follows:

$$M_1 = \alpha_1(K_{\max} \times J)\log\left(\mathcal{I}/K_{\max}\right) + \beta_1 \tag{16}$$

where $K_{\max}$ is the statistical upper bound of K, α_1 and β_1 are experimental values provided in [36].

Result 2: Successful signal recovery can be guaranteed when the number of sensed channels is not less than M, which is defined as follows:

$$M = \alpha_2\left(K \times J\right)\log\left(\mathcal{I}/K\right) + \beta_2 \tag{17}$$

where α_2 and β_2 are also experimental values provided in [36].

Algorithm 2 Proposed Shrink Algorithm

Input: $\mathbf{P}^{\mathbf{EC}}$, Δ_1, K_{ini}, $Iter_{\max}$, and λ_r.
Initialization: $l = 1$, $K^l_{\max} = K_{\mathrm{ini}}$, $M^l_2 < 0$.
while $M^l_2 < 0$ or $l \leq Iter_{\max}$ **do**
  Calculate M^l_1 with $K^l_{\max}$ by (16);
  Calculate $\hat{K}^l$ by (14) and (15) with M^l_1 channels;
  Calculate M^l with $\hat{K}^l$ by (17);
  Calculate $M^l_2 = M^l - M^l_1$;
  Update $K^{l+1}_{\max} = K^l_{\max} - \Delta_1$ and $l = l + 1$.
end while
return $K^l_{\max}$, and $\hat{K}^l$.

Normally, the historical spectrum occupancy $K_{\max}$ is used as a statistical upper bound of rank K, which is utilized to determine the minimal sampling rates at SUs in order to guarantee exact signal recovery. In practice, spectrum occupancies are dynamic, which makes the statistical value $K_{\max}$ not suitable for the practical spectrum occupancy. One possible scenario is that $K_{\max}$ is much larger than K, which leads to that the number of data values collected for the rank estimation being much more than that for exact signal recovery. As a result, it is a waste of data acquisition costs at SUs. Another scenario is that $K_{\max}$ is much smaller than K, which leads to inexact signal recovery as the number of collected data is insufficient for successful signal recovery. In order to obtain the exact rank of the matrix at the FC, we propose a rank estimation algorithm to get $K_{\max}$ adaptively.

In the proposed rank estimation algorithm, for the scenario that $K_{\max}$ is much larger than K, a shrink algorithm is proposed to update $K_{\max}$, as shown in Algorithm 2. In the l-th iteration, M^l_1 is calculated by (16) with $K^l_{\max}$. Additionally, the complete matrix can be obtained by (14) with the M^l_1 number of sensed channels, and then $\hat{K}^l$ can be determined by (15). Additionally, the number of sensed channels M^l required for exact recovery is obtained by (17). Subsequently, M^l_2 is updated as $M^l_2 = M^l - M^l_1$, and $K^{l+1}_{\max}$ is obtained by deducting the step size Δ_1 from $K^l_{\max}$. This iterative process is repeated until $M^l_2 > 0$ or the maximal iteration number $Iter_{\max}$ is reached. The outputs of Algorithm 2 are the updated $K_{\max}$ and $\hat{K}^l$.

Because of the over-utilizing of Algorithm 2 caused by unsuitable step size Δ_1, output $K_{\max}$ from Algorithm 2 can become smaller than the real rank K. This leads to an M_1 that is insufficient for the exact rank estimation $\hat{K}$. As illustrated in Algorithm 3, an enlargement algorithm is proposed to enlarge $K_{\max}$ until exact rank estimation can be achieved. The outputs of Algorithm 2, $K_{\max}$ and $\hat{K}^l$, are taken as inputs of Algorithm 3. In the l-th iteration, $K^{l+1}_{\max}$ is updated by adding step size Δ_2 to $K^l_{\max}$. Subsequently, the updated M^l_1 is determined by (16) to achieve exact rank estimation. Additionally, $\hat{K}^l$ is determined by (14) and (15). This iterative process continues until the difference between $K^l_{\max}$ and $\hat{K}^l$ becomes less than the error tolerance ε or the maximal iteration number $Iter_{\max}$ is reached. The updated $K_{\max}$ is taken as the output of the proposed rank order estimation, and used for the rank input to Algorithm 1.

Algorithm 3 Proposed Enlargement Algorithm

Input: $\mathbf{P}^{\mathbf{EC}}$, Δ_2, $K_{\max}$, $\hat{K}^l$, λ_r, $Iter_{\max}$ and ε.
Initialization: $l=1$, $\hat{K}^l_{\max}=K_{\max}$.
1: **while** $\left|\hat{K}^l - K^l_{\max}\right| \le \varepsilon$ or $l \le Iter_{\max}$ **do**
2: Update $K^{l+1}_{\max}=K^l_{\max}+\Delta_2$ and $l=l+1$;
3: Calculate M^l_1 with $K^l_{\max}$ by (16);
4: Calculate $\hat{K}^l$ by (14) and (15) with M^l_1 channels.
5: **end while**
6: **return** $\hat{K}^l=K^l_{\max}$.

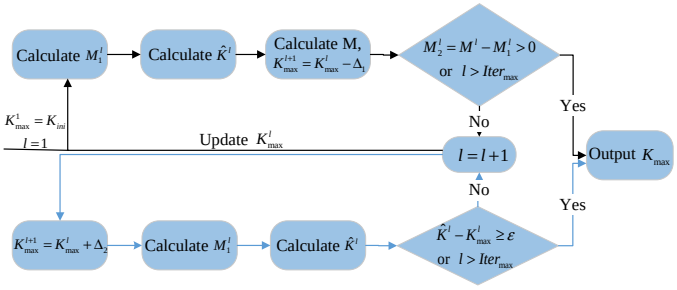


Fig. 3: Procedure of the proposed rank estimation algorithm.

The whole process of the proposed rank estimation algorithm is illustrated in Fig. 3. Once the algorithm starts with the inputs, the updated $K^l_{\max}$ and $\hat{K}^l$ obtained by (15) are returned as the outputs of Algorithm 2. Subsequently, Algorithm 3 is triggered if the difference between $\hat{K}^l$ and $K^l_{\max}$ is no less than the error tolerance ε. The $\hat{K}^l$ is returned until the stop condition is satisfied or the maximal iteration number $Iter_{\max}$ is reached.

C. Estimation Strategy on Number of Corrupted Channels

As illustrated in Algorithm 1, the number of corrupted channels L collected at the FC is required as one of the inputs. However, L is usually unknown and needs to be estimated. Therefore, an estimation of the number of channels corrupted by malicious users L is required. However, if the estimated number of corrupted channels $\hat{L}$ is smaller or greater than its real value L, the performance of Algorithm 1 will be degraded significantly. More specifically, if $\hat{L}$ is underestimated, some of the corrupted channels will still be used to perform MC at the FC, which will definitely result in recovery errors in the reconstructed matrix. If $\hat{L}$ is overestimated, some of the uncorrupted channels will be removed during the MC process, and therefore the MC process would result in more than one solution, as no enough uncorrupted measurements are available for MC. Consequently, exact MC is difficult to achieve as the number of available uncorrupted channels may be insufficient.

As proved in [30], a sufficient condition for the non-uniqueness of a matrix $\mathbf{P}^{\mathbf{\Omega}}$ is given as follows: suppose the number of sensed channels is $\hat{M}$, and they are randomly distributed among the complete matrix $\mathbf{P}^{\mathbf{\Omega}} \in \mathcal{C}^{\mathcal{I}\times J}$. Let us define ΔL as the difference between the overestimated number of corrupted channels $\hat{L}$ and the real number of corrupted channels L. If $\Delta L > \frac{\left(\hat{M}-L\right)}{\max(\mathcal{I},J)} - K > 0$, then the reconstructed matrix is non-unique.

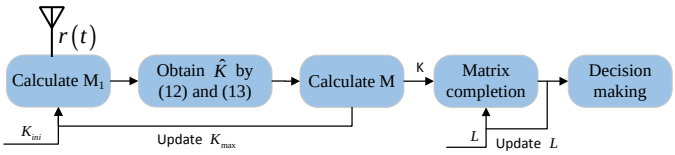


Fig. 4: Whole procedure of the proposed malicious user detection framework based on low-rank matrix completion.

In the proposed framework, an initial guess L_0 for the number of channels corrupted by malicious users is used as one input for Algorithm 1. It is an iterative process to update the number of corrupted channels in combining with Algorithm 1. In each iteration, after Algorithm 1 is performed, the value of the $\hat{L}$th largest term in set $\boldsymbol{S}_{\mathbf{\Omega}} = \left\{\left(\left(\mathbf{P}^{\mathbf{\Omega}}\right)_{ij} - p^{\mathbf{EC}}_{ij}\right)^2, (i,j)\in\mathbf{\Omega}\right\}$ should be checked. If it is less than a predefined tolerance l_{tol}, $\hat{L}$ is determined to be overestimated, and some of the removed channels are uncorrupted. The numerical analyses in [30] has proven that ΔL can be bounded by $(l_{\min}-K)$, where $l_{\min}$ is defined as the minimum number of sensed channels in one row or one column of the incomplete matrix with $\left(\hat{M}-L\right)$ elements at the FC. More specifically, let us define $\hat{l}_{\min}$ as the minimal number of sensed channels in one row or one column of the incomplete matrix with $\left(\hat{M}-\hat{L}\right)$ elements. If $\hat{l}_{\min}$ is less than the rank K, the estimated number of corrupted channels $\hat{L}$ is updated as $\hat{L}=\hat{L}+\hat{l}_{\min}-K$. If $\hat{l}_{\min}$ is no less than K and $\left(p_{ij}-p^{\mathbf{EC}}_{ij}\right)^2$ is less than Γ, the exact matrix is reconstructed. Consequently, the iterative process for MC is terminated. Otherwise, if $\hat{l}_{\min}$ is greater than K and $\left(p_{ij}-p^{\mathbf{EC}}_{ij}\right)^2$ is greater than Γ, $\hat{L}$ is considered to be underestimated. As a result, the value of $\hat{L}$ should be updated to be $\rho_1\hat{L}$, where $\rho_1>1$ is a properly selected constant. The updated $\hat{L}$ is then taken as one input for the MC process in the next iteration until the exact matrix is recovered or the iteration number reaches its upper bound $I_{\max}$.

With the proposed rank estimation algorithm and the number of corrupted channel estimation algorithm described above, the whole procedure of the proposed malicious user detection framework with low-rank MC can be summarized as shown in Fig. 4. Within this framework, it is worth noting that the exact recovery cannot be achieved if the number of sensed channels $\hat{M}$ is smaller than M_1. In this case, the FC should coordinate SUs in the CSS network to sense more channels until $\hat{M}\le M_1$.

D. Data Acquisition Costs and Computational Complexity

If the whole spectrum of interest is entirely sensed by the set of SUs at each location, the number of sensed channels obtained at the FC is $\hat{M}=\mathcal{I}\times J$. In such as case, a complete matrix is available at the FC. Otherwise, if only part of the whole spectrum is sensed, only an incomplete matrix can be constructed at the FC with $\hat{M}<\mathcal{I}\times J$. Herein, the compression ratio can also be defined as $\gamma=\frac{\hat{M}}{\mathcal{I}\times J}$, which is the ratio of the number of the sensed channels $\hat{M}$ in the

incomplete matrix $\mathbf{P^E}$ to the total number of channels $\mathcal{I} \times J$ in the complete matrix $\mathbf{P^\Omega}$ at the FC.

Without invoking the MC technique, the total number of channels to be sensed at J different locations in the CSS network is $\hat{M}_1 = \mathcal{I} \times J$, regardless of the presence of malicious users. Additionally, considering a MC based CSS network without malicious users, in order to make sure the complete matrix can be recovered exactly at the FC, the minimal number of channels to be sensed is

$$\hat{M}_2 = \gamma_{\min} \times \mathcal{I} \times J \tag{18}$$

where $\gamma_{\min}$ is the lower bound of compression ratio to guarantee exact MC, which is dependent on the rank K of the matrix to be recovered and the MC solvers. Furthermore, for the CSS network with the presence of malicious users considered in this paper, with invoking **Algorithm 1**, the minimal number of channels to be sensed is

$$\hat{M}_3 = \hat{\gamma}_{\min} \times \mathcal{I} \times J \tag{19}$$

where $\hat{\gamma}_{\min} \in [\gamma_{\min}, 1]$ is the minimal compression ratio that can be achieved by **Algorithm 1**. The exact value of $\hat{\gamma}_{\min}$ is dependent on the rank K and the channel corruption ratio κ.

Here, the channel corruption ratio $\kappa = \frac{L}{\mathcal{I} \times J}$ is defined as the ratio of the number of corrupted channels L to the total number of channels $(\mathcal{I} \times J)$ to be sensed in the considered CSS network. When there is no malicious user in the CSS network, i.e., $\kappa = 0$, we have $\hat{M}_2 < \hat{M}_1$. With the presence of malicious users, i.e., $\kappa > 0$, we have $\hat{M}_3 \geq \hat{M}_2$. Meanwhile, $\hat{M}_3 \leq \hat{M}_1$ according to the definition of $\hat{M}_3$ in (19). Therefore, no matter whether malicious users exist in the CSS network, if MC is invoked at the FC, the number of channels that needs to be sensed is less than the case without invoking MC technique. As a result, the data acquisition costs can be significantly reduced by the proposed malicious user detection framework, in comparison with the conventional malicious user detection method.

Complexity of the proposed malicious user detection framework lies in two parts: i) rank estimation; and ii) malicious user detection. For the rank estimation part, the computational complexity is $\mathrm{O}\left(\mathcal{I}^3 J^3\right)$, which comes from solving the l_1-norm minimization. For the malicious user detection part, the computational complexity mainly comes from step 2 in Algorithm 1. In each iteration of Algorithm 1, the computational complexity of solving step 2 is $\mathrm{O}\left(\left(\hat{M} + \mathcal{I} + J\right) K^2 + \mathcal{I} K^3\right)$. As a result, the complexity of Algorithm 1 is bounded by $\mathrm{O}\left(I_{\max}\left(\left(\hat{M} + \mathcal{I} + J\right) K^2 + \mathcal{I} K^3\right)\right)$. Due to the low-rank property of the matrix at the FC, the value of matrix rank, K, is usually small, which leads to an acceptable complexity.

IV. Numerical Analyses

In this section, numerical analyses of the proposed malicious user detection framework are presented. The proposed framework is tested on both simulated and real-world TVWS signals.

In the simulation, the total number of channels is $\mathcal{I} = 40$, which is the number of TVWS channels in the UK. Each channel is either fully occupied by a PU or vacant. The locations of PUs are randomly distributed among the spectrum of interest. Here, the *size* of a CSS network refers to the number of different locations J where SUs are implemented in that CSS network. In simulation settings, the size of a CSS network changes from small scale ($J = 1$) to large scale ($J = 400$). The power values of corrupted channels $\tilde{p}_{ij}$ are assumed to be random within the range of $[p^{\mathbf{E}}_{\min}, p^{\mathbf{E}}_{\max}]$. Without considering channel fading, $p^{\mathbf{E}}_{\max}$ is normalized as 1 which refers to the power values of occupied channels, and $p^{\mathbf{E}}_{\min}$ is 0 indicating that the channel is vacant. In the simulation, with considering the channel fading and different compression ratios, the values of $p^{\mathbf{E}}_{\max}$ and $p^{\mathbf{E}}_{\min}$ can be variable in each trial. The maximal number of iterations for Algorithm 1 is $I_{\max} = \mathit{Iter}_{\max} = 200$. The step size is $\Delta_1 = \Delta_2 = 1$, and $\varepsilon = 1$.

A. Numerical Results Using Simulated Signals

In this subsection, simulation results of the proposed framework are illustrated by using the simulated signals.

1) Results of Proposed Rank Estimation Algorithm: The saved sampling costs, which refer to the saved number of channels to be sensed for guaranteeing exact MC, is shown in Fig. 5 with increasing number of sensing periods and dynamic spectrum occupancies. In this scenario, varying spectrum occupancies result in changing the rank K of the matrix at the FC. Therefore, the upper bound $K_{\max}$ of the rank should change accordingly. The rank change points are marked by circle in Fig. 5. Here, the two-step compressive sensing based spectrum sensing algorithm (TS-CS-SS) [36] utilizing fixed $K_{\max}$ is illustrated as the benchmark. The same initial value for $K_{\max}$ is used for both TS-CS-SS and the proposed rank estimation algorithm. Therefore, the sampling costs saved by these two algorithms are same at the initial sensing period in Fig. 5. When the sparsity of the spectral signals changes, the saved sampling costs caused by TS-CS-SS algorithm keep constant as $K_{\max}$ is fixed. However, exact MC cannot be guaranteed if $K_{\max}$ is much smaller than K, or extra data acquisition costs are caused if $K_{\max}$ is much larger than K. For the proposed rank estimation algorithm, the saved sampling costs is degraded at each change point in order to guarantee exact MC. However, the saved sampling costs are increased gradually after the sparsity change point. The proposed rank estimation algorithm guarantees exact MC in dynamic spectrum occupancy environment, with the extra burden on sampling in comparison with the TS-CS-SS algorithm. In practice, spectrum occupancy is assumed to be the stable within several periods. Therefore, the proposed rank estimation algorithm is reliable for practical scenario, as it can adjust $K_{\max}$ to be closer to the real rank K after a few periods of the K change point. More accurate estimation of K makes the proposed algorithm outperform TS-CS-SS in terms of the saved sampling costs with guarantee on exact MC.

2) Results of the Case with Unknown Number of Corrupted Channels: In Fig. 6, the influences of the incorrect estimation on the number of corrupted channels $\hat{L} = \rho L$ are analyzed. Here, ρ, named as estimation accuracy ratio, is defined as the ratio of the estimated number of corrupted channels $\hat{L}$

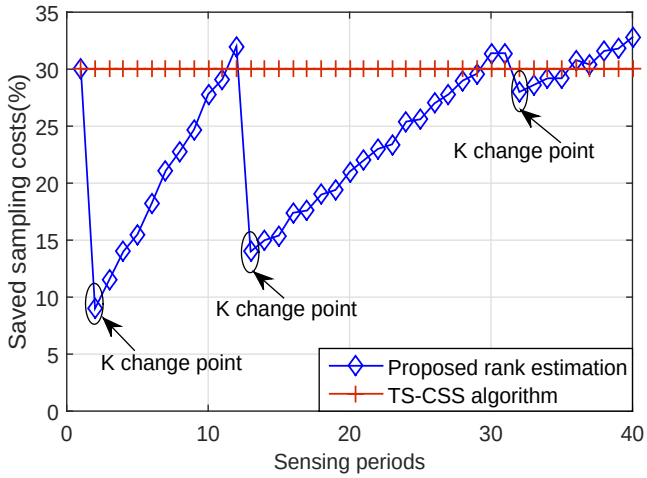


Fig. 5: Saved sampling costs with varying spectrum occupancies.

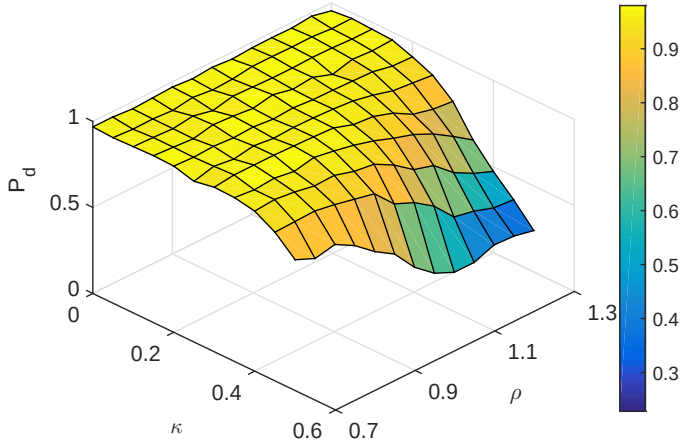


Fig. 6: Detection probability P_d of the proposed malicious user detection framework with different estimation accuracy ratios ρ and channel corruption ratios κ. The compression ratio is $\gamma = 100\%$, and network size is $J = 40$.

to the actual number of corrupted channels L. In this case, we choose $J = 40$ and $K = 1$ to simplify the simulation process. Additionally, the estimated number of corrupted channels $\hat{L} = \rho L$ varies from $0.7L$ to $1.3L$. It is shown in Fig. 6 that detection probability P_d of the proposed malicious user detection framework gets degraded when the estimated number of corrupted channels $\hat{L}$ is overestimated, i.e. $\rho > 1$, especially in the case with high level of channel corruption ratio, i.e. $\kappa \geq 0.5$. It is further noted that P_d would only be degraded slightly if the number of corrupted channels $\hat{L}$ is underestimated. In the following simulations, by invoking the proposed estimation strategy for the number of corrupted channels, the correct estimation of the corrupted channels $\hat{L}$ is taken as one input of **Algorithm 1**.

3) Results of the Proposed Malicious User Detection Framework: Fig. 7 shows the verification of our proposed method for its effectiveness of dealing with malicious users.

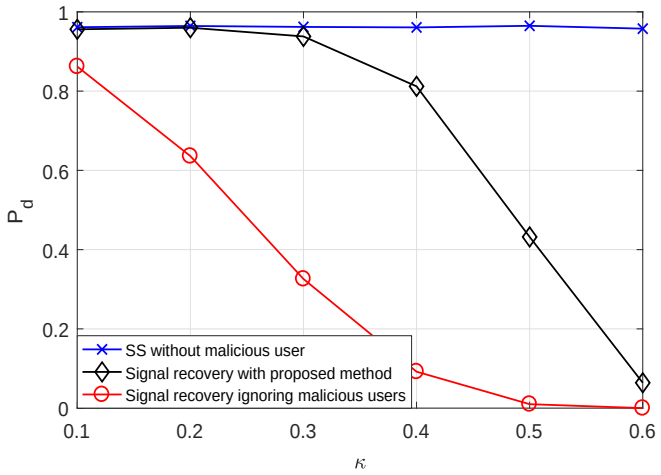


Fig. 7: Detection performance comparison between the proposed method and various benchmarks versus malicious user ratio κ. Compression ratio is $\gamma = 1.0$, matrix rank is $K = 1$, and network size is $J = 40$.

Here, we provide curves for the following four approaches:

- Approach 1: spectrum sensing without malicious users, labelled by '$\times$' in the figure.
- Approach 2: with the existence of malicious users, signal recovery is performed by solving the nuclear norm optimization problem as adopted in [25] for the multiple nodes case. It is labelled by '$\diamond$' in Fig. 7.
- Approach 3: with the existence of malicious users, our proposed malicious user detection method is adopted to remove malicious users during matrix recovery process. It is labelled by '$\circ$' in Fig. 7.

In this scenario, the number of active PUs in the spectrum of interest is 1, which results in the rank of matrix at the FC as $K = 1$. Compression ratio is $\gamma = 1.0$ to avoid any performance degradation caused by insufficient number of collected measurements at the FC. By comparing the curves for approach 1 with approach 2, we can observe that the detection performance is degraded when malicious users attack the system. This is because that the collected measurements used for signal recovery at the FC are corrupted by malicious users, which leads to inaccurate signal recovery. With invoking our proposed method, i.e., approach 3, it can be noted that the detection performance is improved significantly in comparison with approach 2, which validates the effectiveness of our proposed framework. It is also worth noting that the detection performance degrades when more channels are corrupted by malicious users.

In order to further illustrate how malicious users influence the signal recovery accuracy, we also provide a comparison on the relative recovery error achieved by approach 2 and approach 3 as specified above. Here, the relative recovery error is defined as $\frac{\left\|\hat{P}^{\Omega} - P^{\Omega}\right\|_2^2}{\left\|P^{\Omega}\right\|_2^2}$. As shown in Fig. 8, we can see that the relative recovery error achieved by the traditional compressive spectrum sensing approach, i.e., approach 2, is much higher than our proposed approach, i.e., approach 3.

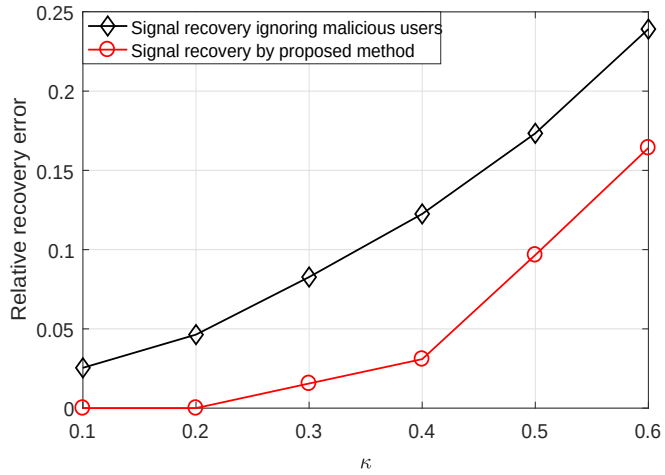


Fig. 8: Recovery accuracy comparison between the proposed method and traditional compressive spectrum sensing approach versus malicious user ratio κ. Compression ratio is $\gamma = 1.0$, matrix rank is $K = 1$, and network size is $J = 40$.

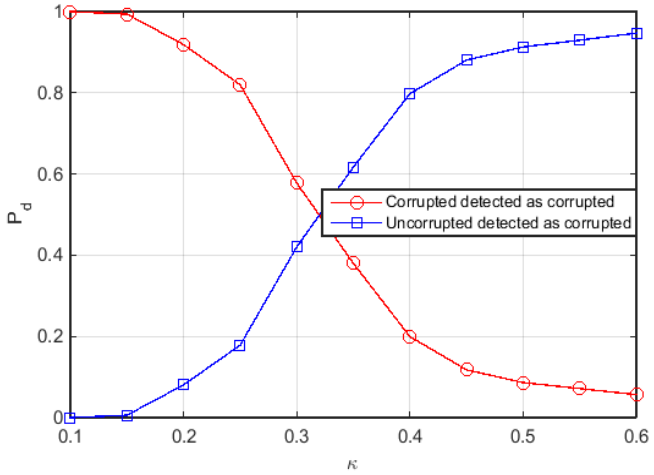


Fig. 9: Corrupted channel detection probability versus malicious user ratio κ. The compression ratio is $\gamma = 1.0$, matrix rank is $K = 1$, and network size is $J = 40$.

This is caused by the existence of malicious users. When these corrupted values are included in the measurements to perform signal recovery at the FC, the recovery error increases. The more channels are corrupted by malicious users, the worse recovery accuracy can be achieved. When our proposed malicious user detection approach is adopted to remove those corrupted measurements during the process of recovery ordinal matrix, the recovery accuracy is increased significantly. This result matches with the detection performance trend shown in Fig. 7.

Fig. 9 plots the probabilities of detecting the corrupted channels as corrupted (P_d) and that of incorrectly classifying uncorrupted channels as corrupted (P_f) versus channel corruption ratio κ. In order to eliminate any possible influence caused by an insufficient number of measurements at the FC, the compression ratio is $\gamma = 1.0$ in this case. From the figure, we can observe that the probability of incorrectly classifying the uncorrupted channels increases with higher κ. The reason is that with higher κ, the number of uncorrupted power values collected at the FC is reduced, which results in inexact MC. It is also worth noting that probability that the corrupted channels are successfully removed during the MC process decreases with increasing κ, which is caused by the inexact MC at the FC as well. We can further note that as more uncorrupted channels are determined as corrupted and removed during the MC process, the number of uncorrupted values used for the MC is further reduced. As a result, the recovery performance of MC degrades.

Fig. 10 plots the detection probability P_d of the proposed malicious user detection framework versus channel corruption ratio κ. It is worth noting that the detection probability P_d of the proposed malicious user detection framework decreases when more channels are corrupted by malicious users in the considered networks. Specifically, we can observe that when channel corruption ratio is increased to 0.6, the detection probability P_d is heavily degraded regardless of the compression ratio. It is reasonable as the number of uncorrupted values is insufficient to guarantee exact MC at the FC.

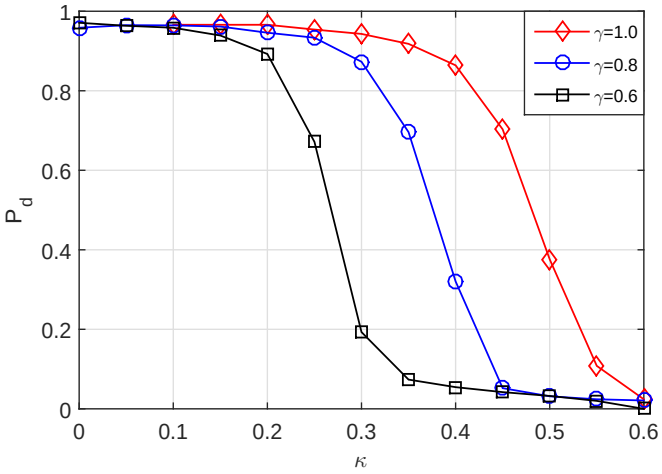


Fig. 10: Detection probability P_d of the proposed malicious user detection framework with different channel corruption ratios κ. The compression ratio γ is 0.6, 0.8, and 1.0, respectively. Here, network size is $J = 40$.

Fig. 11 illustrates how the detection probability P_d of the proposed malicious user detection framework varies against channel corruption ratio κ under different network sizes J. With a fixed κ, the number of corrupted channels L increases with larger network size J, as the channel corruption ratio is defined as $\kappa = \frac{L}{I \times J}$ and I is fixed to be 40. Additionally, as labelled with black ovals in Fig. 11, for the two cases with $J = 200$ and $J = 400$, if κ is 0.4 and 0.2, respectively, the number of corrupted channels becomes the same $L = 3200$. Similarly, as labelled with black ovals in Fig. 11, for the cases with $J = 40$ and $J = 200$, if κ is set to 0.5 and 0.1, the number of corrupted channels $L = 800$ becomes the same. In these

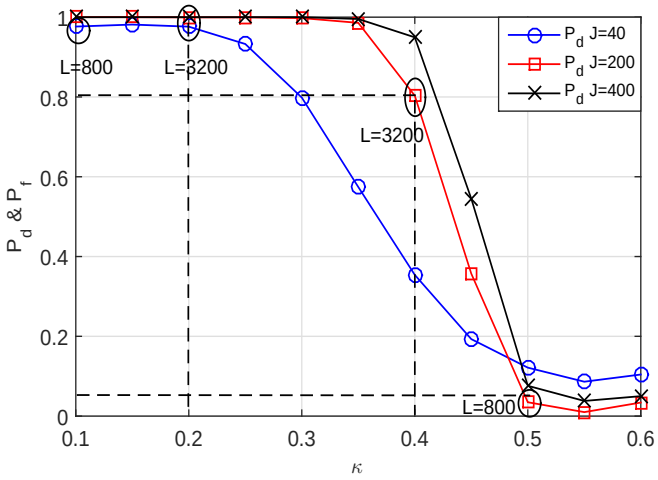


Fig. 11: Detection probability P_d of the proposed malicious user detection framework with different channel corruption ratios κ. The network size J is 40, 200, and 400, respectively. The compression ratio is $\gamma = 1.0$, and matrix rank is $K = 4$.

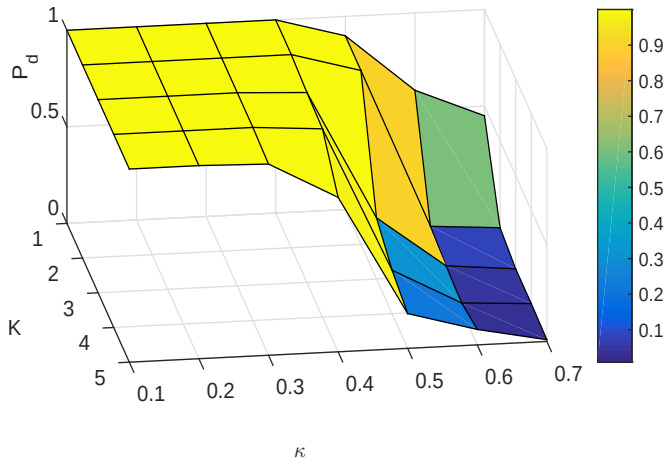


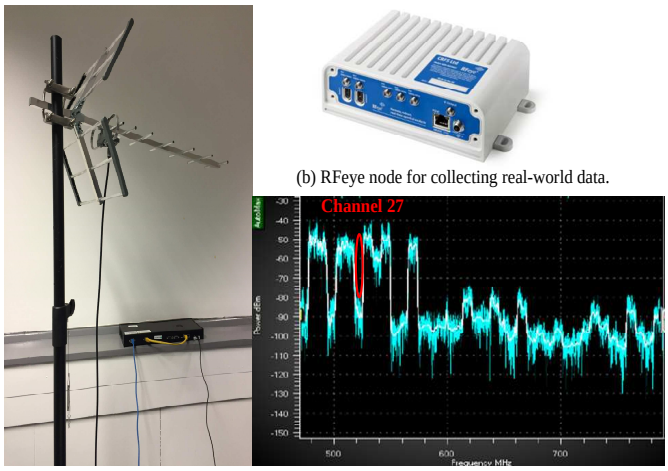
Fig. 12: Detection probability P_d of the proposed malicious user detection framework with different matrix ranks K and different channel corruption ratios κ. The compression ratio is $\gamma = 1.0$, and network size is $J = 400$.

two considered scenarios, we notice that detection probability P_d of the CSS network with large size is higher than that with smaller size when the number of corrupted channels is fixed. This supports the proposition that the higher spatial diversity of the CSS network (i.e., larger CSS network size), the better defense against the same number of corrupted channels.

In Fig. 12, the detection probability P_d of the proposed malicious user detection framework is presented with different channel corruption ratios κ and different ranks K. The rank K of the matrix is determined by the number of active PUs in the spectrum of interest. The positions of the active PUs are randomly generated in the spectrum of interest. In this case, the compression ratio γ is set to 1.0 to avoid any possible performance degradation caused by an insufficient number of sensed channels at the FC. We can see that the detection performance improves with decreasing rank K of the matrix as well as decreasing channel corruption ratio κ. This observation is reasonable, as exact MC requires more observed measurement at the FC with increasing rank K and channel corruption ratio κ.

B. Numerical Results Using the Real-World Signals

The UK regulator Ofcom has conducted a series of trials on the TVWS pilots [37, 38]. One of the trials has been carried out at Queen Mary University of London. There are $I = 40$ channels over TVWS in total, ranging from 470 MHz to 790 MHz. Each TVWS channel is with bandwidth of 8 MHz. Among these TVWS channels, channel 27 is generally vacant, whose frequency ranges from 518 MHz to 526 MHz. During the measurement, channel 27 was randomly corrupted by Digital Video Broadcasting-Terrestrial (DVB-T) signals, which are generated and transmitted temporarily by the setup shown in Fig. 13(a). Real-world signals over TVWS are collected by a portable RFeye node as shown in Fig 13(b). As shown in Fig 13(c), there are 8 channels being occupied by PUs,



(a) Setup for DVB-T signal generation. (b) RFeye node for collecting real-world data. (c) Real-world signal received at the receiver.

Fig. 13: Measurement setup for collecting real-world data at Queen Mary University of London.

including channel 22, 23, 25, 26, 28-30, and 33. When the trial signal is transmitted, as labelled by red circle in Fig. 1, channel 27 became occupied. Due to the limited number of sensing nodes and restricted licence for accessing TVWS, channel 27 was corrupted randomly among different time slots in order to simulate the random distribution of corrupted values among the matrix at the FC. In the following, the proposed malicious user detection framework is tested by real-world signals collected during the trial.

In this case, the same channel is sensed by SUs collected at $J = 50$ different time slots. Malicious users appear in channel 27 randomly among the 50 time slots. Fig. 14 shows the detection probability P_d and the false alarm probability P_f of the proposed malicious user detection algorithm with varying compression ratios γ. The detection performance comparison is demonstrated for the cases with and without malicious users

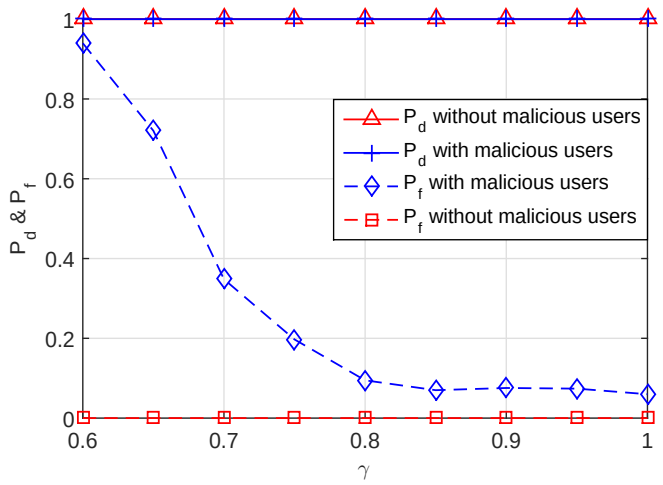


Fig. 14: Detection probability P_d and false alarm probability P_f of the proposed malicious user detection algorithm using real-world signals with different compression ratios γ, and $K = 9$.

in the CSS network. When there are no malicious users in the CSS network, we can see that the perfect detection performance (P_d=1.0 and P_f=0.0) can be achieved by choosing the suitable threshold according to (12) for decision making. If malicious users show up in the CSS network, the false alarm probability P_f becomes higher than the case without malicious users, because a false alarm happens if the corrupted value on channel 27 over TVWS is not removed properly during MC process. It is further noticed that the false alarm probability P_f does not reach 0 when there are malicious users in the CSS network. This is caused by the assumption in the proposed framework that the corrupted channels are assumed to be sparsely and randomly distributed among the sensed channels collected at the FC. This assumption is reasonable and practical as malicious users can show up anywhere and interpolate any channel. However, in the trial, only channel 27 among the 40 TV channels can be corrupted randomly due to the license limitation. The slight behaviour difference of malicious users in the real-world signals causes inexact MC, by utilizing the proposed framework for malicious user detection. Thus, as shown in Fig. 14, the higher P_f is found in comparison with the benchmark (the case without malicious users). However, we can still observe that P_f gets closer to the benchmark with the increasing compression ratio γ, as the recovery error becomes smaller when higher number of sensed channels are available at the FC. This trend matches the simulation results presented in Fig. 10.

V. Conclusions

In this paper, we proposed a low-rank matrix completion (MC) based malicious user detection framework for the secure cooperative spectrum sensing (CSS) networks. As each SU only sensed part of the whole spectrum, the costs of data acquisition at SUs were reduced significantly. Meanwhile, a malicious user detection algorithm was proposed by adopting the adaptive outlier pursuit (AOP) algorithm, in which the channels corrupted by malicious users were removed during the MC process. Additionally, a rank estimation algorithm and an estimation strategy for the number of corrupted channels were proposed to provide inputs for the proposed malicious user detection framework, in order to make it completely blind. Furthermore, the proposed malicious user detection framework was tested on both simulated signals and real-world signals over TV white space spectrum. Numerical analyses showed that the proposed framework alleviated the influences of malicious users with lower data acquisition cost at each individual SU. It can be concluded that the proposed malicious user detection framework is a strong candidate for the secure CSS networks.

Notations

The key notations in this paper are summarized in Table I.

TABLE I: Key notations and definitions

Notations	Definitions
B_j	Number of SUs implemented at the j-th location in a CSS network to sense the spectrum of interest
$\mathcal{I}$	Number of channels among the whole spectrum of interest
J	Number of locations for SUs to sense the same channel in a CSS network (CSS network size)
K	Number of occupied channels (rank of the matrix at the FC)
$K_{\max}$	Statistical upper bound of rank K
$\hat{K}$	Estimated rank of the matrix constructed at the FC
L	Number of corrupted channels collected at the FC
$\hat{L}$	Estimated number of corrupted channels collected at the FC
M	Minimal number of sensed channels for guaranteeing exact MC
M_1	Minimal number of sensed channels for guaranteeing exact rank estimation
$\hat{M}$	Number of sensed channels collected at the FC
$\mathbf{\Omega}$	Index set of the complete matrix constructed at the FC
$\mathbf{E}$	Index set for the sensed channels and their locations
$\mathbf{O}$	Index set for the the uncorrupted sensed channels and their locations
$\mathbf{P}^{\mathbf{\Omega}}$	Complete matrix constructed at the FC
$\boldsymbol{P}^{\mathbf{\Omega}}$	Reconstructed matrix for rank order estimation
$\hat{\boldsymbol{P}}^{\mathbf{\Omega}}$	Reconstructed matrix at the FC
$\mathbf{P}^{\mathbf{E}}$	Incomplete matrix with corrupted values collected at the FC
$\mathbf{P}^{\mathbf{EC}}$	Partly corrupted incomplete matrix collected at the FC
p_{ij}	Uncorrupted power value of the i-th channel sensed by the SU at the j-th location
$\tilde{p}_{ij}$	Corrupted power value of the i-th channel sensed by the SU at the j-th location
$\bar{p}_{ij}$	Power value recovered by the rank order estimation
$\hat{p}_{ij}$	Power value recovered by the proposed malicious user detection framework
$\mathbf{\Lambda}$	Binary matrix denoting the uncorrupted channels by one and the corrupted ones by zero
ρ	Estimation accuracy ratio, defined as the ratio of the estimated number of corrupted channels $\hat{L}$ to the real number of corrupted channels L
γ	Compression ratio, defined as the ratio of the sensed channels $\hat{M}$ to the total number of channels to be sensed $\mathcal{I} \times J$ in the complete matrix
κ	Channel corruption ratio, defined as the ratio of the number of corrupted channels L to the total number of channels to be sensed $\mathcal{I} \times J$ in the complete matrix

Acknowledgement

Yue Gao is supported by funding from Physical Sciences Research Council (EPSRC) in the U.K. with Grant

No. EP/L024241/1. Mark D. Plumbley is partly supported by funding from the European Union's Seventh Framework Programme (FP7-PEOPLE-2013-ITN) under grant agreement 607290 SpaRTaN, and the H2020 Framework Programme (H2020-MSCA-ITN-2014) under grant agreement 642685 MacSeNet.

REFERENCES

[1] J. Mitola and G. Q. Maguire, "Cognitive radio: Making software radios more personal," *IEEE Pers. Commun.*, vol. 6, no. 4, pp. 13 –18, Aug. 1999.

[2] A. Ghasemi and E. Sousa, "Collaborative spectrum sensing for opportunistic access in fading environments," in *Proc. IEEE Intl. Symp. New Frontiers Dynamic Spectrum Access Netw. (DySPAN)*, Baltimore, MD, Nov. 2005, pp. 131–136.

[3] I. F. Akyildiz, B. F. Lo, and R. Balakrishnan, "Cooperative spectrum sensing in cognitive radio networks: A survey," *Physical Commun.*, vol. 4, no. 1, pp. 40–62, Mar. 2011.

[4] R. Chen, J. M. Park, Y. T. Hou, and J. H. Reed, "Toward secure distributed spectrum sensing in cognitive radio networks," *IEEE Commun. Mag.*, vol. 46, no. 4, pp. 50–55, Apr. 2008.

[5] G. Ding, Q. Wu, Y.-D. Yao, J. Wang, and Y. Chen, "Kernel-based learning for statistical signal processing in cognitive radio networks: Theoretical foundations, example applications, and future directions," *IEEE Signal Process. Mag.*, vol. 30, no. 4, pp. 126–136, Jun. 2013.

[6] Q. Yan, M. Li, T. Jiang, W. Lou, and Y. Hou, "Vulnerability and protection for distributed consensus-based spectrum sensing in cognitive radio networks," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 900–908.

[7] L. Zhang, G. Ding, Q. Wu, Y. Zou, Z. Han, and J. Wang, "Byzantine attack and defense in cognitive radio networks: A survey," *IEEE Commun. Surveys Tutorials*, vol. 17, no. 3, pp. 1342–1363, Apr. 2015.

[8] P. Kolodzy and I. Avoidance, "Spectrum policy task force," *Federal Commun. Comm., Washington, DC, Rep. ET Docket*, no. 02-135, Jun. 2002.

[9] UK Office of Communications (Ofcom), "Statement on Cognitive Access to Interleaved Spectrum" Jul. 2009.

[10] Z. Tian and G. Giannakis, "Compressed sensing for wideband cognitive radios," in *Proc. IEEE Intl. Conf. Acoustics Speech Signal Process.(ICASSP)*, Honolulu, HI, Apr. 2007, pp. 1357–1360.

[11] Y. Wang, Z. Tian, and C. Feng, "Collecting detection diversity and complexity gains in cooperative spectrum sensing," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, pp. 2876 –2883, Aug. 2012.

[12] H. Li, "Reconstructing spectrum occupancies for wideband cognitive radio networks: A matrix completion via belief propagation," in *Proc. IEEE Intl Conf. Commun. (ICC)*, Cape Town, South Africa, May 2010, pp. 1–6.

[13] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations Comput. Math.*, vol. 9, no. 6, pp. 717–772, Dec. 2009.

[14] W. Wang, H. Li, Y. Sun, and Z. Han, "Securing collaborative spectrum sensing against untrustworthy secondary users in cognitive radio networks," *EURASIP J. Adv. Signal Process.*, vol. 2010, pp. 1–15, Jan. 2010.

[15] R. Chen, J.-M. Park, and K. Bian, "Robust distributed spectrum sensing in cognitive radio networks," in *Proc. IEEE INFOCOM*, Phoenix, AZ, Apr. 2008, pp. 13–18.

[16] P. Kaligineedi, M. Khabbazian, and V. K. Bhargava, "Malicious user detection in a cognitive radio cooperative sensing system," *IEEE Trans. Wireless Commun.*, vol. 9, no. 8, pp. 2488–2497, Jun. 2010.

[17] A. W. Min, K. G. Shin, and X. Hu, "Secure cooperative sensing in IEEE 802.22 WRANs using shadow fading correlation," *IEEE Trans. Mobile Comput.*, vol. 10, no. 10, pp. 1434–1447, Oct. 2011.

[18] S. Kalamkar, A. Banerjee, and A. Roychowdhury, "Malicious user suppression for cooperative spectrum sensing in cognitive radio networks using Dixon's outlier detection method," in *Proc. National Conf. Commun. (NCC)*, Kharagpur, Feb. 2012, pp. 1–5.

[19] A. W. Min, K. H. Kim, and K. G. Shin, "Robust cooperative sensing via state estimation in cognitive radio networks," in *IEEE Symp. on New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, May 2011, pp. 185–196.

[20] F. Penna, Y. Sun, L. Dolecek, and D. Cabric, "Detecting and counteracting statistical attacks in cooperative spectrum sensing," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 1806–1822, Apr. 2012.

[21] L. Duan, A. W. Min, J. Huang, and K. G. Shin, "Attack prevention for collaborative spectrum sensing in cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 9, pp. 1658–1665, Oct. 2012.

[22] H. Li and Z. Han, "Catch me if you can: An abnormality detection approach for collaborative spectrum sensing in cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3554–3565, Nov. 2010.

[23] W. Wang, L. Chen, K. Shin, and L. Duan, "Thwarting intelligent malicious behaviors in cooperative spectrum sensing," *IEEE Trans. Mobile Comput.*, vol. 14, no. 11, pp. 2392–2405, Nov. 2015.

[24] J. Meng, W. Yin, H. Li, E. Hossain, and Z. Han, "Collaborative spectrum sensing from sparse observations in cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 2, pp. 327–337, Feb. 2011.

[25] Z. Qin, Y. Gao, M. D. Plumbley, and C. G. Parini, "Wideband spectrum sensing on real-time signals at sub-Nyquist sampling rates in single and cooperative multiple nodes," *IEEE Trans. Signal Process.*, vol. 64, no. 12, pp. 3106–3117, Jun. 2016.

[26] Z. Qin, Y. Liu, Y. Gao, M. Elkashlan, and A. Nallanathan, "Wireless powered cognitive radio networks with compressive sensing and matrix completion," *IEEE Trans. Commun.*, vol. 65, no. 4, pp. 1464–1476, Apr. 2017.

[27] Y. Ma, Y. Gao, Y. C. Liang, and S. Cui, "Reliable and efficient sub-Nyquist wideband spectrum sensing in cooperative cognitive radio networks," vol. 34, no. 10, pp. 2750–2762, Oct. 2016.

[28] Z. Qin, Y. Gao, and C. G. Parini, "Data-assisted low complexity compressive spectrum sensing on real-time signals under sub-nyquist rate," *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1174–1185, Feb. 2016.

[29] Z. Qin, Y. Gao, M. Plumbley, C. Parini, and L. Cuthbert, "Low-rank matrix completion based malicious user detection in cooperative spectrum sensing," in *Proc. IEEE Global Conf. Signal Info. Process. (GlobalSIP)*, Austin, TX, Dec. 2013, pp. 1186–1189.

[30] M. Yan, Y. Yang, and S. Osher, "Exact low-rank matrix completion from sparsely corrupted entries via adaptive outlier pursuit," *J. Sci. Comput.*, vol. 56, no. 3, pp. 433–449, Sep. 2013.

[31] ——, "Robust 1-bit compressive sensing using adaptive outlier pursuit," *IEEE Trans. Signal Process.*, vol. 60, no. 7, pp. 3868–3875, Jul. 2012.

[32] Y. C. Liang, Y. Zeng, E. C. Y. Peh, and A. T. Hoang, "Sensing-throughput tradeoff for cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 4, pp. 1326–1337, Apr. 2008.

[33] N. Boumal and P.-A. Absil, "RTRMC: A Riemannian trust-region method for low-rank matrix completion," in *Proc. Advances Neural Info. Process. Systems 24 (NIPS)*, Granada, Spain, Dec. 2011, pp. 406–414.

[34] J. Tropp, J. Laska, M. Duarte, J. Romberg, and R. Baraniuk, "Beyond nyquist: Efficient sampling of sparse bandlimited signals," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 520–544, Jan. 2010.

[35] E. Candes, "Compressive sampling," in *Proc. Intl. Congress Math.*, vol. 3, Madrid, Spain, Aug. 2006, pp. 1433–1452.

[36] Y. Wang, Z. Tian, and C. Feng, "Sparsity order estimation and its application in compressive spectrum sensing for cognitive radios," *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, pp. 2116–2125, May 2012.

[37] *UK Office of Communications (Ofcom), "Ofcom TV White Spaces Pilot Update Event"*, Jun. 2014. [Online]. Available: http://stakeholders.ofcom.org.uk/binaries/spectrum/whitespaces/TVWS_Pilot_Update_event_26-06-14_presentation_PUBLISH.pdf

[38] O. Holland, S. Ping, A. Aijaz, J. M. Chareau, P. Chawdhry, Y. Gao, Z. Qin, and H. Kokkinen, "To white space or not to white space: That is the trial within the Ofcom TV white spaces pilot," in *Proc. IEEE Intl. Symp. Dynamic Spectrum Access Netw. (DySPAN)*, Stockholm, Sweden, Sep. 2015, pp. 11–22.

Zhijin Qin (S'13-M'16) received her Bachelor degrees from Beijing University of Posts and Telecommunications in China in 2012, and her Ph.D degree from Queen Mary University of London in UK in 2016. She has joint at Lancaster University in UK as a Lecturer (Assistant Professor) in the School of Computing and Communications since August 2017. Before that, she was with Imperial College London as a Research Associate. She won the best paper award at Wireless Technology Symposium 2012.

Her research interests include low power wide area network in Internet of Things, compressive sensing and machine learning in wireless communications, and nonorthogonal multiple access. She currently serves as an Editor of IEEE ACCESS. She has served as a TPC member for many IEEE conferences such as GLOBECOM'16, ICC'16, VTC'15 and VTC'14.

Yue Gao (S'03-M'07-SM'13) is a Reader in Antennas and Signal Processing, and Director of Whitespace Machine Communication Lab in the School of Electronic Engineering and Computer Science at Queen Mary University of London (QMUL) in the UK. He worked as Research Assistant, Lecturer (Assistant Professor) and Senior Lecturer (Associate Professor) at QMUL after having received his PhD degree from QMUL in 2007. He is currently leading a team developing theoretical research into practice in the interdisciplinary area among smart antennas, signal processing, spectrum sharing and internet of things (IoT) applications. He has published over 130 peer-reviewed journal and conference papers, 2 patents, and 2 book chapters. He is a co-recipient of the EU Horizon Prize Award on Collaborative Spectrum Sharing in 2016, and Research Performance Award from Faulty of Science and Engineering at QMUL in 2017.

He is an Editor for the IEEE Transactions on Vehicular Technology, IEEE Wireless Communication Letter and China Communications. He is serving as Cognitive Radio Symposium Co-Chair of the IEEE GLOBECOM 2017. He has served as the Signal Processing for Communications Symposium Co-Chair for IEEE ICCC 2016, Publicity Co-Chair for IEEE GLOBECOM 2016, and General Chair of the IEEE WoWMoM and iWEM 2017. He is a Senior Member of IEEE, a Secretary of the IEEE Technical Committee on Cognitive Networks, and an IEEE Distinguished Lecturer of Vehicular Technology Society.

Mark D. Plumbley (S'88-M'90-SM'12-F'15) received the B.A.(Hons.) degree in electrical sciences and the Ph.D. degree in neural networks from University of Cambridge, Cambridge, U.K., in 1984 and 1991, respectively. From 1991 to 2001, he was a Lecturer with Kings College London, London, U.K., before moving to Queen Mary University of London, London, U.K. in 2002, later becoming Director of the Centre for Digital Music. In 2015, he joined the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, U.K., as Professor of Signal Processing. His main research interest is the analysis and processing of audio and music signals, using techniques such as matrix factorization, sparse representations, and deep learning. He is a Member of the IEEE Signal Processing Society Technical Committee on Signal Processing Theory and Methods.